Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 48 (2015) 606 – 611

Procedia Computer Science

International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2015)

Conference Organized by Interscience Institute of Management and Technology,

Bhubaneswar, Odisha, India

An Active Learning Framework for Human Hand Sign Gestures and Handling Movement Epenthesis Using Enhanced Level Building Approach

Elakkiya R[a], Selvamani K[b]

[a]*Research Scholar, Department of Computer Science & Engineering, Anna University, Chennai-600025, India*
[b]*Assistant Professor, Department of Computer Science & Engineering, Anna University, Chennai-600025, India*

Abstract

Human hand detection and segmentation plays an important role in sign language recognition and human machine interaction. In this paper, a novel approach for learning a vision-based hand detection system is introduced. The main contribution of this paper includes robust boosting-based framework for real-time detection of a hand in unconstrained and heterogeneous environments. The proposed system makes use of efficient representative features which allows fast computation while changing the hand appearances and background. Moreover, this proposed strategy efficiently improves the performance while reducing the effort of hand labeling. Experimental results show that the proposed method is practically more flattering as it meets the requirements of real-time performance, accuracy and robustness. This system has been proved to work well with a reasonable amount of training samples and was computationally found to be more effective and efficient.

Peer-review under responsibility of scientific committee of International Conference on Computer, Communication and Convergence (ICCC 2015)

Keywords: Sign Gestures; Movement Epenthesis; Level Building Algorithm; Machine Learning;

Peer-review under responsibility of scientific committee of International Conference on Computer, Communication and Convergence (ICCC 2015)
doi:10.1016/j.procs.2015.04.142

1. Introduction

Human computer interface (HCI) is the greatest challenge in sign language recognition and it offers a unique opportunity for the development of motion recognition algorithms. The problem of detecting human hand has many applications in gesture recognition, sign language, human computer interactions, robot control, etc. Hand detection and recognition have long been an active research area in computer vision based systems. There exist two major approaches to obtain desired information about hand's appearance. One reliable approach is to either use a glove with sensors or to use a camera with markers on the hand. This approach is expensive and very obtrusive. The other approach is a computer vision based system. This approach has become more favorable as powerful machine learning methods come to deal with complex nature of hand appearances, changing environment, and computational efficiency.

Besides, most of the systems mentioned above used only simple features such as a Haar-like wavelet to represent the object of interest. This Haar-like feature alone is not powerful tool to represent complex objects like hands. Moreover, the procedure for training a boosted classifier is straight forward. There is no attempt to utilize the availability of current training classifier for increasingly improving performance. In this work, an active learning approach for learning a hand detection system is proposed, which tackles the difficulties stipulated above.

In this research paper, a novel sign recognition strategy that does not require explicit modeling of movement epenthesis is presented. This approach that does not use explicit movement epenthesis model. Gestures base of signs to be recognized by the proposed system not the model of movement epenthesis. The classical Level Building algorithm (CLB) is enhanced to meet the above criteria as well as to match the continuous sign sentences without explicit movement epenthesis models. The proposed approach outlined a hand segmentation based on key frames and it adopts a histogram based representation as features to detect the sign gestures. This system utilizes both motion and skin cues to recognize gestures in cluttered background and it also uses a set of key frames to model the background. The detection algorithm exploits the fact that the hand is changing its orientation faster than other parts during signing. To create the model base during training, it needs the sign key frames in continuous sentences without considering its associated movement epenthesis.

The main focus is to detect a human hand in an unconstrained environment. The output is an image region containing a detected hand, which will facilitate further application processes such as hand gesture recognition or tracking. This paper is organized in the following structure. Section 2 gives a brief review of the related work. Section 3 presents our proposed approach for building the framework for hand-detection system. Section 4 dedicates our experiments and results. Finally this proposed work concludes with further discussions on future capabilities.

2. Related work

Adaboost learning algorithms have been proved to be one of the most powerful approaches, in terms of speed and accuracy, for visual object learning and detection. The spirit of boosting is a powerful ensemble learning algorithm, which combines a number of weak learners to produce a strong classifier with high accuracy. Recently, there has been considerable interest on computer vision problems in applying boosting based techniques with impressive success. Some examples are: on-line co-training of Javed et al. [12], and the selection of discriminative features of Grabner and Bischof [11]. These methods use the same underlying boosting algorithm proposed by Oza et.al [17], minimizes the classification error while updating the weak classifiers. Despite of the success of boosting methods in a wide range of vision based problems, such as face detection [21] and the detection of persons [14], the application of boosting-based learning for hand detection and recognition has limitations. In [14], a hand detection system is developed based on boosted classifiers proposed in [13]. The system is trained to detect just six well-defined hand gestures from still images. A similar approach is employed in [6]. Experiments were done in a dynamic environment, but only six defined gestures were concerned and hands appeared in large at the center of an image. Another work[13], a single classifier is employed but limited to perform only precise gesture, i.e, detector can be accomplished with about 15 degrees in-plane rotations. Very recently, [3] proposed a real-time hand gesture recognition system, which is also based on the standard Viola & Jones system. In our approach, we use efficient integral image representation for fast calculation of hands features. The features include Haar wavelet [9], local

orientation histogram [4] and a simplified version of local binary patterns [16], which can be fast computed on integral images.

A novel version of Adaboost to train the detector is used. The algorithm performs on-line updating on the ensembles of features during the training process. By on-line interactive training, the classifier is updated as a new sample is provided, and therefore we can reduce effort for labeling of training samples. In addition, a strategy to exploit the availability of classifier is proposed during training time to automatically label good samples for learning meanwhile increasingly improve performance. So, labeling effort is further reduced. Besides, the system allows to access and update on the seen data to prevent drifting of classifier due to vast changes of hand object over time. The developed framework results in a robust and automatic hand detection system and achieves a high performance. The system is flexible since we do not use any constraint to model the hand as well as background environment, or motion information.

3. Proposed hand detection system

The proposed approach to build our hand detection system is presented in figure 2 followed by sub-sections, with the ideas on efficient representation of skin colour model, the boosting methods which will be used for feature selection, active training process which will allows efficient learning and the Nested Dynamic Programming Using Enhanced Level Building Algorithm.

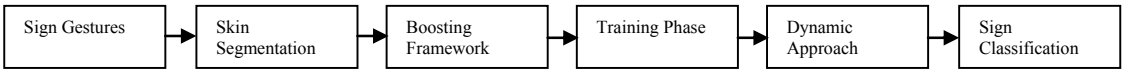

Fig. 2. Proposed Hand Detection System

3.1. Skin Colour Model

The purpose of the initial hand segmentation is to collect training samples for the SVM classifier, which is implemented using skin model by defining a fixed colour range in one colour space using histograms. For Support Vector Machine, given a linear separable training set {x1, x2, ...:,xi, ..., xn} with labels {y1, y2, ...,yi, ...,yn} where yi €{-1, 1}, an optimal binary classification is yielded for the skin data by solving an optimisation problem. For a given continuous gesture video, the skin colour is extracted on the first several frames so that the training set containing skin pixels and non-skin pixels could be obtained. The obtained training set of the SVM Classifier is constructed and used to segment the frames one by one in exploitation.

3.2. Boosting Algorithm

Boosting methods works by combining a number of weak learning algorithms into a strong one. Many variations of boosting have been proposed, e.g. Real-Boost [10], LP-Boost [5]. We focus on the discrete AdaBoost (adaptive boosting) algorithm which is already presented in [22]. The probability p(x) is updated such that it increases for the samples that are misclassified. The corresponding weight is decreased if the sample is classified correctly. Therefore, the algorithm focuses on the difficult examples. At each boosting iteration a new weak classifier is added and the process is repeated until a certain stopping condition is met (e.g. a given number of weak classifiers are trained). Finally, a strong classifier hstrong(x) is computed as linear combination of a set of N weak classifiers hweak n (x). Boosting also has been developed for feature selection. The main idea is that each feature corresponds to a single weak classifier and boosting selects an informative subset from these features.

3.3. Training and detection

In a learning version, the training process is performed by iteratively labeling samples from the images and updating parameters for the model. The labeled samples can be positive or negative. In order to minimize the hand labeling effort, an active learning strategy is applied. The key idea used is that the user has to label only examples

which are not correctly classified by the current classifier. The classifier is evaluated and updated after each labeling of a sample. The new updated classifier is applied again on the same image or on a new image, and the process continues. Since labeling of samples in the training phase is an interactive process with human supervision, we can intuitively choose to label the most informative and discriminative sample at each update, which allows the parameters of the model to be updated in a greedy manner with respect to minimizing the detection error. It also avoids labeling redundant samples that do not contribute to the current decision boundary.

A classifier is a system that inputs (typically) a vector of discrete and/or continuous feature values and outputs a single discrete value. A learner inputs a training set of examples (xi, yi), where xi = (xi, 1 . . . xi,d) is an observed input and yi is the corresponding output, and outputs a classifier. The test of the learner is whether this classifier produces the correct output y for future examples xt. Because the classifier is adaptive to newly coming samples, it may forget some sample that it has learned so far. This can be referred to as drifting or over-adaptiveness of the classifier.

3.4. Nested Dynamic Programming Using Enhanced Level Building Algorithm

One naive way to obtain the solution is to enumerate among all the possible sign sequence candidates Si , compute the warping distance score between SI and T, find the Si with minimum score. Clearly the computational complexity of such an approach is prohibitive. Hence, we adopt an sequential approach to build this optimal sign sequence using a framework called Level Building and enhance it to allow for movement epenthesis labels. Each level corresponds to the possible order of signs or movement epenthesis in the test sentence. Thus, the first level is concerned with the first possible label in the sentence, and so on. Each level is associated with a set of possible start and end locations within the sequence. And at each level we store the best possible match for each combination of end point from the previous level. The optimal sequence of signs and movement epenthesis labels is constructed by backtracking. For each level l, we store the optimal cost for matching between sign Si and with the ending frame as m using a 3 dimensional array A.

Tmj denotes a subsequence of the test sequence that starts at the jth frame and ends at the mth frame. Hence Ail(m) gives us the minimum cumulative score for matching the ith model sign, Si to the test sequence upto them-th frame, for the lth sign label in the sequence. The choice of the cost for labeling a frame as movement epenthesis is a crucial one. We choose this by considering the distribution of match and non-match scores between signs in the training set. A match score is defined to the cost of matching different instances of the same sign and a non-match score is cost of matching instances of different signs. These scores are computed using dynamic warping and using the same frame to frame distance function used in the Level Building algorithm. They are normalized by the length of the warping path.

3.5. Sign Representation

Since the major contribution of this work is the enhanced Level Building algorithm, the low-level representation used for completeness is sketched. Since our test is done based on pure video data, a segmentation scheme to segment the hands out of the scene to form the feature vectors for each frame is developed. This step is automatic, but has some noise. The assumption that we make is that the hands move faster than other objects in the scene (including the face), and that the hand area can be somewhat localized by skin color detection. We used the mixed Gaussian model, we use a safe threshold such that non skin pixels can be falsely classified as skin pixels. We represent the possibly changing (but slowly) background, using a set of key frames. These key frames are identified as frames that are sufficiently different from each other. We sequentially search for them, starting from the first frame, which is always chosen to be a key frame. We compute the difference of any frame with previous key frame. If the non-component size in the threshold difference image is large then the frame is labeled as the next key frame. This process continues until the end of the sequence. Then we compute the difference image of each frame to the key frames.

4. Experimental Results

We have conducted extensive experimentation of the approach in the context of the task of recognizing continuous American Sign Language sentences from image sequences. We present not only visual results of labeling continuous ASL sentences, but also quantify the performance. We compare the performance with theclassical Level Building, which does not account for movement epenthesis. In the results, empirical evidence of the optimality of the choice of the α parameter is used to decide on the me mapping cost is presented and the impact of the grammar model is also presented.

4.1. Dataset

The vocabulary consists of signs that a deaf person would need to communicate with security personnel at airports. The video data is taken at 30 fps, with an image resolution of 460 by 290. There are 39 different signs that are articulated in 25 different continuous sentences. (Note that for approaches that explicitly model me we would need around 1000 sentences to capture the variations between signs.) Some signs appear more than once in some sentences. The total number of individual sign instances in the dataset is 73. There are 5 instances of each sentence. Some sentences have significant variations between multiple instances of the same sentence.

4.2. Results

A labeling result for three sentences is diagrammatically presented in Fig 4. Each horizontal bar represents a sentence and is partitioned into signs or me blocks. The size of each block is proportional to the number of frames corresponding to that label. For each sentence we present the ground truth as determined by an ASL expert and the results from the algorithm. It is obvious that the signer is signing at different speeds for each sign. For instance, the sign I is spread over a large number of frames. Apart from a 1 to 2 frame mismatch at the beginning and the end, the labeling match pretty well. To quantitatively evaluate the results, we use errors as advocated. In first case, different hand's postures in a complex background are detected. The number of labeled samples for training the system together with detection rates to obtain that performance is shown in Table 1. If the recognized sentence inserted i.e a sign that does not actually exist, one insertion error is reported. The recognized sentence reports a wrong sign, it is considered as a substitution error. These errors are computed automatically by computing the Levenshtein distance using a dynamic programming approach between the actual results and manually labeled ground truth. Fig. 5 shows the error rates we obtained with the optima for each test set in the 5-fold validation experimentation.

Table 1. No. of Training samples with Detection Rate

Dataset N.	Frames N.	Samples	Detection rate%
1 (Seq.2)	1976	129	99.6
1 (Seq.3)	600	67	100
2	674	95	100

Fig. 4. Sign level error rates set in the 5-fold cross validation Level experiments with ASL data.

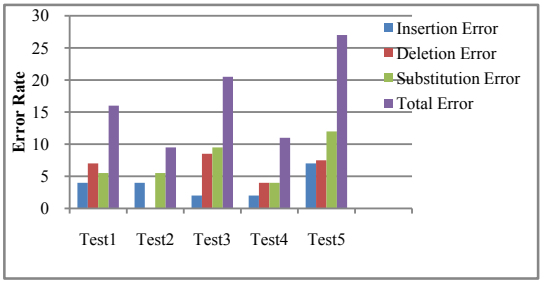

Fig. 5. Comparison of Enhanced Level Building vs classical Building

5. Conclusion

In this research paper, Enhanced Classical Level Building algorithm is proposed to built around dynamic programming to address the problem of movement epenthesis in continuous sign sentences. This approach does not explicitly model movement epenthesis, the demand on annotated training video data is low. The performance of enhanced Level Building with classical Level building algorithm is compared. Our extensive experiments demonstrate the robustness of the matching process to different parameters. The proposed enhanced Level Building algorithm solves the general problem of recognizing motion patterns from stream of compositions of motion patterns with portions without considering explicit model.

References

1. S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part- based representation, Transactions on Pattern Analysis and Machine Intelligence, 2004, 1475–1490
2. C.-C. Chang and C. M. Pengwu. Gesture recognition approach for sign language using curvature scale space and hidden markov model. ICME'04. 2004, 1187–1190.
3. Q. Chen, N. Georganas, and E. Petriu. Real-time vision based hand gesture recognition using haar-like features. *Instrumentation and Measurement Technology Conference Proceedings*, 2007,1-6.
4. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *In Proc., CVPR, volume 1, ,San Diego, CA, USA, IEEE Computer Society*, 2005, 886–893
5. A. Demiriz, K. Bennett, and J. Shawe-Taylor. Linear programming boosting via column generation. Machine Learning, 2002, 46(1-3): 225–254.
6. H. Francke, J. Solar, and R. Verschae. Real-Time Hand Gesture Detection and Recognition Using Boosted Classifiers and Active Learning, *chapter Lect. Notes in Computer Science. Springer Berlin/Heidelberg*, 2007, , 533–547
7. W. T. Freeman and M. Roth. Orientation histograms for hand gesture recognition. *In Intl. Workshop FG'95, IEEE Computer Society*, 1995, 296–301.
8. Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Jour. of Computer and System Sciences*, , 1997, 55(1):119–139.
9. H. Grabner and H. Bischof. On-line boosting and vision. *In Proceedings Conference on Computer Vision and Pattern Recognition*, 2006, volume 1, pages 260–267.
10. O. Javed, S. Ali, and M. Shah. detection and classification of moving objects using progressively improving detectors. *In Proceedings CVPR*, San Diego, CA, USA, 2005. IEEE Computer Society, pages 695–700.
11. M. Kolsch and M. Turk. Analysis of rotational robustness of hand detection with a viola-jones detector. *Proceedings, ICPR*, 2004, 3:107–110 .
12. M. Kolsch and M. Turk. Robust hand detection. *Proceedings, FG04, IEEE*, , 2004, pages 614–619.
13. K. Levi and Y.Weiss. Learning object detection from a small number of examples: The importanceof good features. *In Proceedings of Conference on Computer Vision and Pattern Recognition*, , Washington, DC, USA, 2004, pages 53–60.
14. T. Ojala, M. Pietik¨ainen, and T. M¨aenp¨a¨a. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence,* 2002, 24(7):971–987.
15. E. Ong and R. Bowden. A boosted classifier tree for hand shape detection. *Proc. FG04, IEEE*, 2004, pages 889–894.
16. N. Oza and S. Russell. Experimental comparisons of and batch versions of bagging and boosting. *In Proceedings ACM SIGKDD,* 2001, 359–364.
17. N. Oza and S. Russell. bagging and boosting. In Proceedings Artificial Intelligence and Statistics, 2001, 105–112.
18. J.-H. Park and Y.-K. Choi. On-line learning for active pattern recognition. *IEEE Signal Processing Letters*, 1996, 3(11):301–303.
19. K. Tieu and P. Viola. Boosting image retrieval. *In Proceedings, CVPR,* Hilton Head, SC, USA, 200, pages 228–2350.
20. J. Triesch and C. Malsburg. Robust classification of hand postures against complex backgrounds. *Proceedings FG'96,* 1996, 170–175.
21. C. Wang and C. Wang. Hand posture recognition using adaboost with SIFT for human robot interaction. *In Proceedings ICAR'07, J*eju, Korea, August 2007.
22. Elakkiya R, Selvamani K, Kannan A, "An intelligent framework for recognizing sign language from continuous video sequence using boosted subunits", *Forth International Conference on SEISCON*, 2013,297-304 .